Linear Regression

Stephen B. Holt, Ph.D.

ROCKEFELLER COLLEGE
OF PUBLIC AFFAIRS & POLICY
UNIVERSITY AT ALBANY State University of New York

April 14, 2022

Returning to the Road Map

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question (convenience samples, random samples, stratified or multistage samples).
4. Calculate measures to test hypotheses put forward about the relationship of interest (measures of central tendency, measures of spread, test statistics).
5. Organize and report results (graphs, tables, interpretations of measures).

Focusing on Steps 2 and 4: Hypothesis Testing

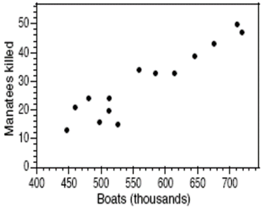
1. State the null and alternative hypotheses and α level of significance
   - Null is a *status quo* assumption about the world you are testing with your sample of data. Stated as $H_o : \mu = X$ where X is an assumption about the true value of μ
   - Alternative is *your* assumption about the world you are testing with your sample of data. Generally, the alternative hypothesis takes the form of $H_1 : \mu \neq X$, $H_1 : \mu > X$, or $H_1 : \mu < X$.
   - α is a probability, from 0 to 1, that represents the maximum threshold of a p-value you will accept for rejecting the null. Conventionally, social scientists use $\alpha = 0.05$.
2. Calculate t-statistic to test the null hypothesis
   - $t = \frac{(\overline{X} - \mu)}{\frac{s}{\sqrt{n}}}$ using the mean, standard deviation, and n from your sample and plugging in your null hypothesis for μ
3. Use the absolute value of t to find the p-value.
4. Compare the p-value to α; if $p < \alpha$, reject the null hypothesis.

Reminders

- Hypothesis testing is always about whether a *statistic* (e.g, $\overline{X}, \overline{X}_1 - \overline{X}_2$) accurately reflects a *parameter of interest* (e.g., $\mu, \mu_1 - \mu_2$).
- A *parameter* can be the value of a single variable in a typical observation in a population OR the typical relationship between two variables in a typical observation in a population.
- The logic of hypothesis testing for a relationship between two variables is very similar to the logic of testing a statistic from a sample - how confident are we that our estimate of the relationship is not due to random chance?

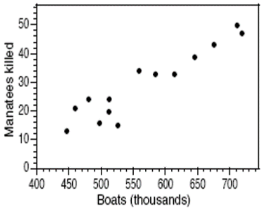
Linear Regression Setup

- Linear regression continues our effort at the same goal we've had in previous weeks: using a sample to estimate a population parameter (thus far, μ) and test hypotheses about the population parameter.



Linear Regression Setup

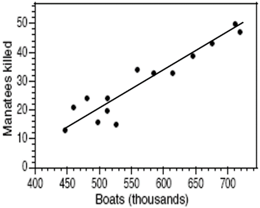
- Linear regression continues our effort at the same goal we've had in previous weeks: using a sample to estimate a population parameter (thus far, μ) and test hypotheses about the population parameter.



- Now we move to a parameter that captures a relationship between two variables in a population, similar to two-sample hypothesis testing. We've seen scatterplots of x and y before. They also come from random samples and change across samples.

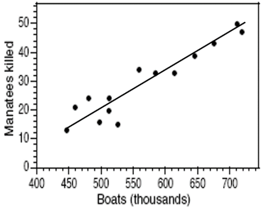
Linear Regression Setup

- In our brave new world, we are still interested in an underlying population parameter, in this case the average outcome Y or μ_y.



Linear Regression Setup

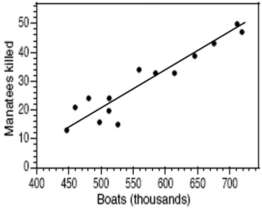
- In our brave new world, we are still interested in an underlying population parameter, in this case the average outcome Y or μ_y.



- Linear regressions, as the name implies, expresses the relationship of x and y as a linear relationship. The goal is to use the line that fits the relationship observed in the data to learn about the population mean response μ_y as a function of our explanatory variable X.

Linear Regression Setup

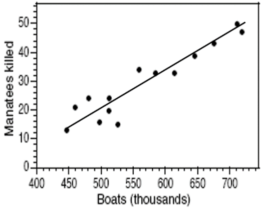
- In our brave new world, we are still interested in an underlying population parameter, in this case the average outcome Y or μ_y.



- Linear regressions, as the name implies, expresses the relationship of x and y as a linear relationship. The goal is to use the line that fits the relationship observed in the data to learn about the population mean response μ_y as a function of our explanatory variable X.
- Mathematically expressed: $\mu_y = \beta_0 + \beta_1 x$

Linear Regression Setup

- In our brave new world, we are still interested in an underlying population parameter, in this case the average outcome Y or μ_y.



- Linear regressions, as the name implies, expresses the relationship of x and y as a linear relationship. The goal is to use the line that fits the relationship observed in the data to learn about the population mean response μ_y as a function of our explanatory variable X.
- Mathematically expressed: $\mu_y = \beta_0 + \beta_1 x$
- We also want to know if β_x, the relationship observed, is statistically significant (i.e., not attributable to chance or sampling error).

Statistical Model for Linear Regression

- In the population, there is a linear regression relationship: $\mu_y = \beta_0 + \beta_1 x$.

Statistical Model for Linear Regression

- In the population, there is a linear regression relationship: $\mu_y = \beta_0 + \beta_1 x$.
- So, because μ_y is some outcome we think is important, like stopping boats from killing manatees, and x can tell us something about what changes μ_y in the population, we collect a sample of data.

Statistical Model for Linear Regression

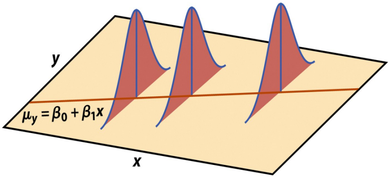
- In the population, there is a linear regression relationship:
  $\mu_y = \beta_0 + \beta_1 x$.
- So, because μ_y is some outcome we think is important, like stopping boats from killing manatees, and x can tell us something about what changes μ_y in the population, we collect a sample of data.
- The sample can then be used to fit the simple model:
  $Data = fit + residual$
  $y_i = (\beta_0 + \beta_1 x) + \varepsilon_i$,
  where ε_i are independent and normally distributed $N(0, \sigma)$.

Statistical Model for Linear Regression

- In the population, there is a linear regression relationship: $\mu_y = \beta_0 + \beta_1 x$.
- So, because μ_y is some outcome we think is important, like stopping boats from killing manatees, and x can tell us something about what changes μ_y in the population, we collect a sample of data.
- The sample can then be used to fit the simple model:
  $Data = fit + residual$
  $y_i = (\beta_0 + \beta_1 x) + \varepsilon_i$,
  where ε_i are independent and normally distributed $N(0, \sigma)$.
- Linear regression assume equal variance of y (i.e., σ is the same for all values of x).

Statistical Model for Linear Regression

- In the population, there is a linear regression relationship: $\mu_y = \beta_0 + \beta_1 x$.
- So, because μ_y is some outcome with think is important, like stopping boats from killing manatees and x can tell us something about what changes μ_y in the population, we collect a sample of data.
- The sample can then be used to fit the simple model: $Data = fit + residual$ $y_i = (\beta_0 + \beta_1 x) + \varepsilon_i$, where ε_i are independent and normally distributed $N(0, \sigma)$.
- Linear regression assume equal variance of y (i.e., σ is the same for all values of x).



Estimating parameters

In the underlying regression model in the population, $\mu_y = \beta_0 + \beta_1 x$, the intercept ($\beta_0$), the slope ($\beta_1$), and the standard deviation of y (σ_y) are all the unknown parameters that we would like to estimate. We rely on the random sample data and least-squares regression to provide unbiased estimates of these parameters (just like with means and two sample tests!).

Estimating parameters

In the underlying regression model in the population, $\mu_y = \beta_0 + \beta_1 x$, the intercept ($\beta_0$), the slope ($\beta_1$), and the standard deviation of y (σ_y) are all the unknown parameters that we would like to estimate. We rely on the random sample data and least-squares regression to provide unbiased estimates of these parameters (just like with means and two sample tests!).

- The value of $\widehat{y}$ from the least-squares regression line is really a prediction of the mean value of y (μ_y) for a given value of x.

Estimating parameters

In the underlying regression model in the population, $\mu_y = \beta_0 + \beta_1 x$, the intercept ($\beta_0$), the slope ($\beta_1$), and the standard deviation of y (σ_y) are all the unknown parameters that we would like to estimate. We rely on the random sample data and least-squares regression to provide unbiased estimates of these parameters (just like with means and two sample tests!).

- The value of $\hat{y}$ from the least-squares regression line is really a prediction of the mean value of y (μ_y) for a given value of x.
- The least-squares regression line ($\hat{y} = b_0 + b_1 x$) is the best estimate of the true population regression line ($\mu_y = \beta_0 + \beta_1 x$)

Estimating parameters

In the underlying regression model in the population, $\mu_y = \beta_0 + \beta_1 x$, the intercept ($\beta_0$), the slope ($\beta_1$), and the standard deviation of y (σ_y) are all the unknown parameters that we would like to estimate. We rely on the random sample data and least-squares regression to provide unbiased estimates of these parameters (just like with means and two sample tests!).

- The value of $\widehat{y}$ from the least-squares regression line is really a prediction of the mean value of y (μ_y) for a given value of x.
- The least-squares regression line ($\widehat{y} = b_0 + b_1 x$) is the best estimate of the true population regression line ($\mu_y = \beta_0 + \beta_1 x$)
  - $\widehat{y}$ unbiased estimate for mean population response μ_y

Estimating parameters

In the underlying regression model in the population, $\mu_y = \beta_0 + \beta_1 x$, the intercept ($\beta_0$), the slope ($\beta_1$), and the standard deviation of y (σ_y) are all the unknown parameters that we would like to estimate. We rely on the random sample data and least-squares regression to provide unbiased estimates of these parameters (just like with means and two sample tests!).

- The value of $\widehat{y}$ from the least-squares regression line is really a prediction of the mean value of y (μ_y) for a given value of x.
- The least-squares regression line ($\widehat{y} = b_0 + b_1 x$) is the best estimate of the true population regression line ($\mu_y = \beta_0 + \beta_1 x$)
    - $\widehat{y}$ unbiased estimate for mean population response μ_y
    - b_0 unbiased estimate for intercept β_0

Estimating parameters

In the underlying regression model in the population, $\mu_y = \beta_0 + \beta_1 x$, the intercept ($\beta_0$), the slope ($\beta_1$), and the standard deviation of y (σ_y) are all the unknown parameters that we would like to estimate. We rely on the random sample data and least-squares regression to provide unbiased estimates of these parameters (just like with means and two sample tests!).

- The value of $\widehat{y}$ from the least-squares regression line is really a prediction of the mean value of y (μ_y) for a given value of x.
- The least-squares regression line ($\widehat{y} = b_0 + b_1 x$) is the best estimate of the true population regression line ($\mu_y = \beta_0 + \beta_1 x$)
  - $\widehat{y}$ unbiased estimate for mean population response μ_y
  - b_0 unbiased estimate for intercept β_0
  - b_1 unbiased estimate for slope β_1

Estimating parameters

Calculating the best fit line ourselves would involve first calculating the slope:

$$\beta_1 = \frac{\sum(x_i - \overline{X})(y_i - \overline{Y})}{\sum(x_i - \overline{X})^2} \tag{1}$$

...and then using the basic form of a line to calculate the intercept:

$$\beta_0 = \overline{Y} - \beta_1 \overline{X} \tag{2}$$

Regression Standard Errors

- Recall that statistical inference for the mean of a sample relies upon an estimate of σ to calculate the standard error ($s.e. = \frac{s}{\sqrt{n}}$, where s is the sample standard deviation). The logic and process is similar for regression estimates.

Regression Standard Errors

- Recall that statistical inference for the mean of a sample relies upon an estimate of σ to calculate the standard error ($s.e. = \frac{s}{\sqrt{n}}$, where s is the sample standard deviation). The logic and process is similar for regression estimates.
- As before, the population standard deviation of y, σ_y, represents the spread of y, only in the population regression model, it reflects the spread of y for each value of x in the population (i.e., the spread of the normal distribution of ε_i around the mean μ_y).

Regression Standard Errors

- Recall that statistical inference for the mean of a sample relies upon an estimate of σ to calculate the standard error ($s.e. = \frac{s}{\sqrt{n}}$, where s is the sample standard deviation). The logic and process is similar for regression estimates.
- As before, the population standard deviation of y, σ_y, represents the spread of y, only in the population regression model, it reflects the spread of y for each value of x in the population (i.e., the spread of the normal distribution of ε_i around the mean μ_y).
- Of course, we don't observe this, but we can use our sample data to compute an estimate of the regression standard error, s, for a sample sized n using the residuals $(y_i - \widehat{y}_i)$:

$$s_{reg} = \sqrt{\frac{\sum residual^2}{n-2}} = \sqrt{\frac{\sum (y_i - \widehat{y}_i)^2}{n-2}} \tag{3}$$

Regression Standard Errors

- Recall that statistical inference for the mean of a sample relies upon an estimate of σ to calculate the standard error ($s.e. = \frac{s}{\sqrt{n}}$, where s is the sample standard deviation). The logic and process is similar for regression estimates.
- As before, the population standard deviation of y, σ_y, represents the spread of y, only in the population regression model, it reflects the spread of y for each value of x in the population (i.e., the spread of the normal distribution of ε_i around the mean μ_y).
- Of course, we don't observe this, but we can use our sample data to compute an estimate of the regression standard error, s, for a sample sized n using the residuals $(y_i - \widehat{y}_i)$:

$$s_{reg} = \sqrt{\frac{\sum residual^2}{n-2}} = \sqrt{\frac{\sum(y_i - \widehat{y}_i)^2}{n-2}} \tag{3}$$

- s provides an unbiased estimate of the regression standard deviation σ, which we can use for inference about the mean population response μ_y.

Regression Standard Errors, continued

The formula is similar for the standard error of the slope (β_1), only the regression standard error (s_{reg}) is divided by the square root of the squared residuals of X:

$$SE_{b1} = \frac{s_{reg}}{\sqrt{\sum(x_i - \overline{x})^2}} \tag{4}$$

Confidence Intervals for Regression Parameters

- Estimating the regression parameters β_0, β_1 is a case of one-sample inference with unknown population variance.

Confidence Intervals for Regression Parameters

- Estimating the regression parameters β_0, β_1 is a case of one-sample inference with unknown population variance.
  - We rely on the t distribution with n-2 degrees of freedom.

Confidence Intervals for Regression Parameters

- Estimating the regression parameters β_0, β_1 is a case of one-sample inference with unknown population variance.
  - We rely on the t distribution with n-2 degrees of freedom.
- A level C confidence interval for the slope (β_1) is proportional to the standard error of the least-squares slope:

$$b_1 \pm t * SE_{b1} \tag{5}$$

Confidence Intervals for Regression Parameters

- Estimating the regression parameters β_0, β_1 is a case of one-sample inference with unknown population variance.
  - We rely on the t distribution with n-2 degrees of freedom.
- A level C confidence interval for the slope (β_1) is proportional to the standard error of the least-squares slope:

$$b_1 \pm t * SE_{b1} \tag{5}$$

- A level C confidence interval for the intercept (β_0) is proportional to the standard error of the least-squares intercept:

$$b_0 \pm t * SE_{b0} \tag{6}$$

Confidence Intervals for Regression Parameters

- Estimating the regression parameters β_0, β_1 is a case of one-sample inference with unknown population variance.
    - We rely on the t distribution with n-2 degrees of freedom.
- A level C confidence interval for the slope (β_1) is proportional to the standard error of the least-squares slope:

$$b_1 \pm t * SE_{b1} \tag{5}$$

- A level C confidence interval for the intercept (β_0) is proportional to the standard error of the least-squares intercept:

$$b_0 \pm t * SE_{b0} \tag{6}$$

- Note that t* is the t-critical value for the t(n-2) distribution with area C between -t* and +t*.

Significance test for the slope

- Once we have calculated the standard error of the least-squares regression line, the process for testing whether the relationship between x and y is statistically significant is analogous to the process for hypothesis testing for a single sample estimate. Here, b_1, or the slope of the least-squares line, is the estimate we use to test a hypothesis about β_1.

Significance test for the slope

- Once we have calculated the standard error of the least-squares regression line, the process for testing whether the relationship between x and y is statistically significant is analogous to the process for hypothesis testing for a single sample estimate. Here, b_1, or the slope of the least-squares line, is the estimate we use to test a hypothesis about β_1.
- As usual, we start with the null hypothesis. Here, since we want to know if our observed relationship between x and y in our sample is significant, we use the null hypothesis that there is no relationship. Formally, $H_0 : \beta_1 = 0$. We can test using a 1- or 2-sided alternative hypothesis.

Significance test for the slope

- Once we have calculated the standard error of the least-squares regression line, the process for testing whether the relationship between x and y is statistically significant is analogous to the process for hypothesis testing for a single sample estimate. Here, b_1, or the slope of the least-squares line, is the estimate we use to test a hypothesis about β_1.
- As usual, we start with the null hypothesis. Here, since we want to know if our observed relationship between x and y in our sample is significant, we use the null hypothesis that there is no relationship. Formally, $H_0 : \beta_1 = 0$. We can test using a 1- or 2-sided alternative hypothesis.
- We will again use the t distribution and calculate our t-score using our estimate of the parameter and estimate of the parameter's spread. In this case, $t = \frac{b_1}{SE_{b1}}$.

Significance test for the slope

- Once we have calculated the standard error of the least-squares regression line, the process for testing whether the relationship between x and y is statistically significant is analogous to the process for hypothesis testing for a single sample estimate. Here, b_1, or the slope of the least-squares line, is the estimate we use to test a hypothesis about β_1.
- As usual, we start with the null hypothesis. Here, since we want to know if our observed relationship between x and y in our sample is significant, we use the null hypothesis that there is no relationship. Formally, $H_0 : \beta_1 = 0$. We can test using a 1- or 2-sided alternative hypothesis.
- We will again use the t distribution and calculate our t-score using our estimate of the parameter and estimate of the parameter's spread. In this case, $t = \frac{b_1}{SE_{b1}}$.
- We then use the t distribution of $t(n-2)$ degrees of freedom to find the p-value.

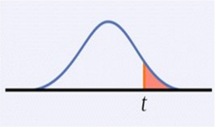
Significance test for the slope

- Once we have calculated the standard error of the least-squares regression line, the process for testing whether the relationship between x and y is statistically significant is analogous to the process for hypothesis testing for a single sample estimate. Here, b_1, or the slope of the least-squares line, is the estimate we use to test a hypothesis about β_1.
- As usual, we start with the null hypothesis. Here, since we want to know if our observed relationship between x and y in our sample is significant, we use the null hypothesis that there is no relationship. Formally, $H_0 : \beta_1 = 0$. We can test using a 1- or 2-sided alternative hypothesis.
- We will again use the t distribution and calculate our t-score using our estimate of the parameter and estimate of the parameter's spread. In this case, $t = \frac{b_1}{SE_{b1}}$.
- We then use the t distribution of $t(n-2)$ degrees of freedom to find the p-value.
- Finally, as before, we compare the p-value to our α threshold and infer whether β_1 is significantly different from 0 given our sample.

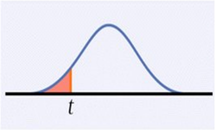
Significance test for the slope

Visually:

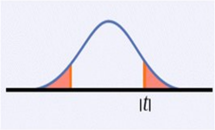
$H_a: \beta_1 > 0$ is $P(T \geq t)$



$H_a: \beta_1 < 0$ is $P(T \leq t)$



$H_a: \beta_1 \neq 0$ is $2P(T \geq |t|)$



Inference for Prediction

- One use for regression is for predicting the value of y, $\widehat{y}$, for any value of x within the range of data tested: $\widehat{y} = b_0 + b_1 x$

Inference for Prediction

- One use for regression is for predicting the value of y, $\widehat{y}$, for any value of x within the range of data tested: $\widehat{y} = b_0 + b_1 x$
- But, just like our estimates $\overline{y}$ from a sample, the regression equation depends on the particular sample drawn. More reliable predictions require inference.

Inference for Prediction

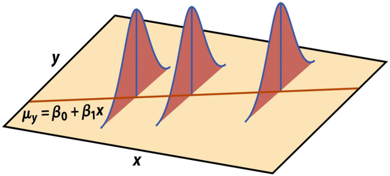
- One use for regression is for predicting the value of y, $\widehat{y}$, for any value of x within the range of data tested: $\widehat{y} = b_0 + b_1 x$
- But, just like our estimates $\overline{y}$ from a sample, the regression equation depends on the particular sample drawn. More reliable predictions require inference.
- To estimate an individual response y for a given value x, we use a prediction interval.

Inference for Prediction

- One use for regression is for predicting the value of y, $\widehat{y}$, for any value of x within the range of data tested: $\widehat{y} = b_0 + b_1 x$
- But, just like our estimates $\overline{y}$ from a sample, the regression equation depends on the particular sample drawn. More reliable predictions require inference.
- To estimate an individual response y for a given value x, we use a prediction interval.
- If we randomly sampled many times, there would be many different values of y obtained for a particular x following a $N(0, \sigma)$ distribution around the mean response μ_y.

Inference for Prediction

- One use for regression is for predicting the value of y, $\widehat{y}$, for any value of x within the range of data tested: $\widehat{y} = b_0 + b_1 x$
- But, just like our estimates $\overline{y}$ from a sample, the regression equation depends on the particular sample drawn. More reliable predictions require inference.
- To estimate an individual response y for a given value x, we use a prediction interval.
- If we randomly sampled many times, there would be many different values of y obtained for a particular x following a $N(0, \sigma)$ distribution around the mean response μ_y.



Confidence Intervals and Prediction

- We can calculate a confidence interval at level C for each predicted value of y, $\widehat{y}$, at each level or value of x.

Confidence Intervals and Prediction

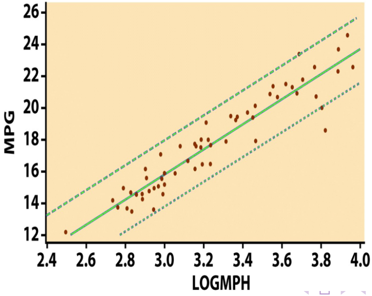
- We can calculate a confidence interval at level C for each predicted value of y, $\widehat{y}$, at each level or value of x.
- The level C prediction interval for a single observation of y when x takes on the value $x*$ is:
  $\widehat{y} \pm t*_{n-2}\, SE_{\widehat{y}}$

Confidence Intervals and Prediction

- We can calculate a confidence interval at level C for each predicted value of y, $\widehat{y}$, at each level or value of x.
- The level C prediction interval for a single observation of y when x takes on the value $x*$ is:
  $\widehat{y} \pm t *_{n-2} SE_{\widehat{y}}$
- The prediction interval represents mainly the error from the normal distribution of the residuals (ε_i).

Confidence Intervals and Prediction

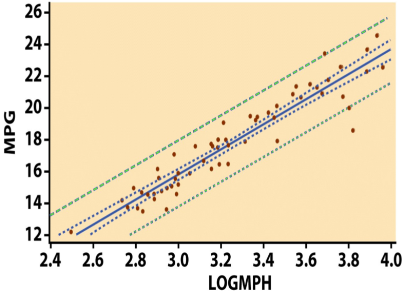
- We can calculate a confidence interval at level C for each predicted value of y, $\widehat{y}$, at each level or value of x.
- The level C prediction interval for a single observation of y when x takes on the value $x*$ is:
  $\widehat{y} \pm t*_{n-2}\, SE_{\widehat{y}}$
- The prediction interval represents mainly the error from the normal distribution of the residuals (ε_i).
  Graphically:



Confidence Intervals for Mean Response (μ_y)

- The confidence interval for μ_y contains, with level C% confidence, the population mean μ_y at a particular level of x.
- The prediction interval contained C% of all the individual values taken by y at a particular value of x.

Graphically:



95% prediction interval for $\hat{y}$ in green
95% confidence interval for μ_y in blue

Coefficient of Determination (R^2)

- The coefficient of determination, generally referred to as R^2 or the square of the correlation coefficient, measures the percentage of the variance in y (vertical scatter from the regression line) that can be explained by changes in x.

Coefficient of Determination (R^2)

- The coefficient of determination, generally referred to as R^2 or the square of the correlation coefficient, measures the percentage of the variance in y (vertical scatter from the regression line) that can be explained by changes in x.
- $R^2 = \dfrac{\textit{variation in y caused by x (the regression line)}}{\textit{total variation in observed y values around the mean}}$

Coefficient of Determination (R^2)

- The coefficient of determination, generally referred to as R^2 or the square of the correlation coefficient, measures the percentage of the variance in y (vertical scatter from the regression line) that can be explained by changes in x.
- $R^2 = \dfrac{\textit{variation in y caused by x (the regression line)}}{\textit{total variation in observed y values around the mean}}$
- More formally:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y}_i)^2} = \frac{SSM}{SST} \tag{7}$$

Attendance